Paleofish Mitochondries

Objectif : Détermination des références reconstructibles

07/02/2024

Suivi du projet

Demande de compte sur :
https://forge-dga.jouy.inra.fr/

Quel work projet ?

Point sur les 89 échantillons

* Au moins 12 FASTQ corrompus (Nb sequences R1/R2 différent → Buffer error)
Listés dans le wiki (accès wiki ?) :

L10_1_BC5_S58
L10_11_SCO_1338_S63
L11_2_SCO_81_S67
L11_6_SCO_116_S69
L2_11_MPHB3_S79
L4_11_MD2_S21
L4_9_MD12_S19
L5_1_MD15_S26
L5_11_MD14_S36
L6_11_SH2000-99-394-g_S46
L6_6_SH2001-106_11-b_S44
L7_14_7013j_S48_R1_001-bonnom.fastq.gz

* 12 FASTQ ne sont pas disponibles dans le répertoire ECOBIOP_ADNa: L1-1, L1-4, L7-15, L8-6, L9-12, L9-8, L9-10, L3-3, L7-11, L7-9, L1-2, L1-7

→ **Une autre source est-elle disponible ? ng6 ? Disque dur ?**

Premiers tests

Un échantillon L2-6 TP1 France Roquemissou (Aveyron) Néolithique (-6000 -2200).

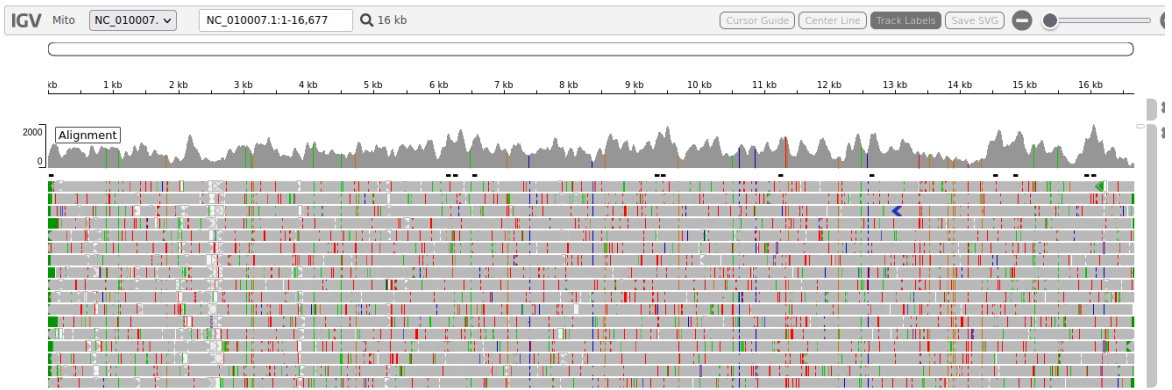
```
samtoobash-4.4$ samtools idxstats L2_6_TP1_S75.sort.bam
NC_010007.1     16677   196145  1487
NC_025589.1     16751   3696    53
NC_025648.1     16737   2240    17
NC_028593.1     16736   1969    21
NC_030175.1     16739   2086    29
NC_001960.1     16665   10558   97
*       0       0       44930
```

→ mapDamage inutile pour trouver les vrais variants car la profondeur est telle que les dommages faits à l'ADN sont masqués.

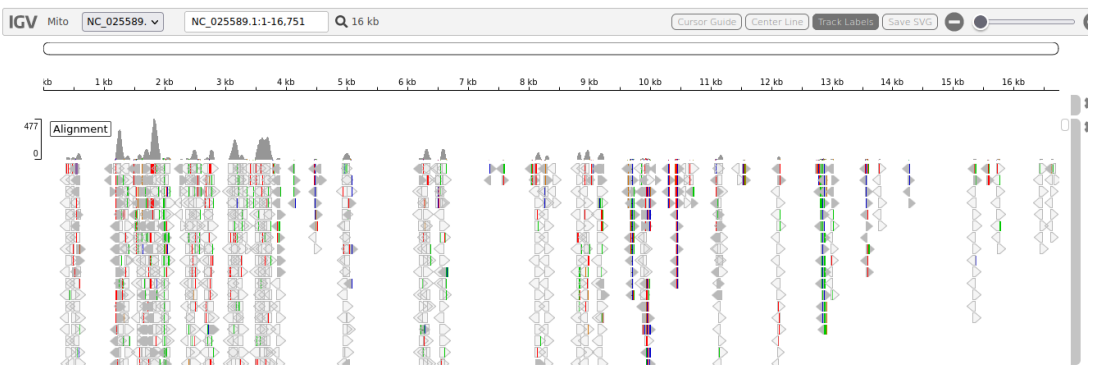
Premiers tests

http://genoweb.toulouse.inra.fr/~sigenae/sarah/mito/igvjs_multifasta.html

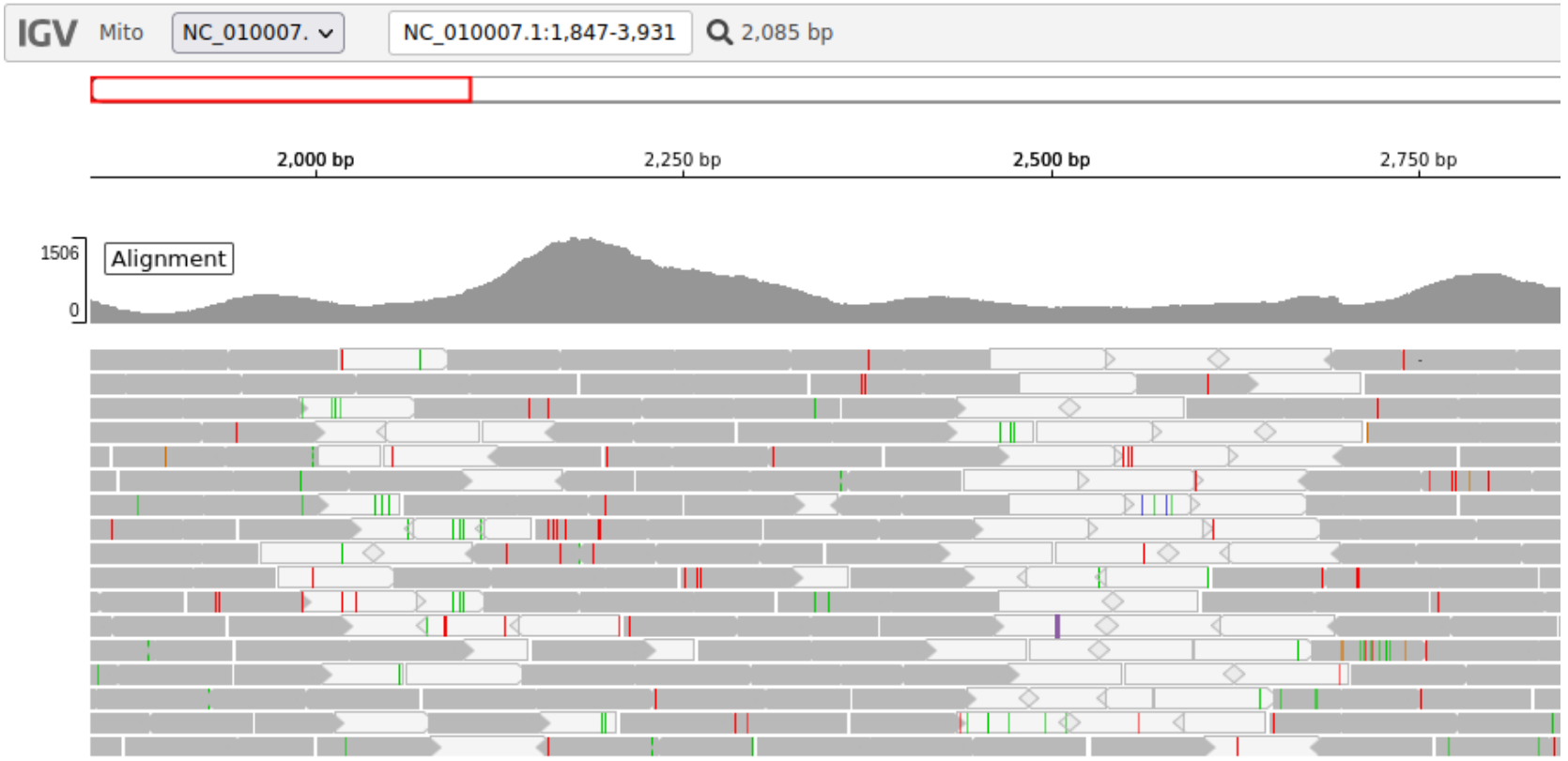
Multi align L2_6_TP1_S75



Multi align L2_6_TP1_S75



Premiers tests : Bonne qualité du consensus



Alignement sur multiFasta

TODO :
- Graph Qte ADN * Nb seq.
- Histo Nb seq. / éch.
- Histo Nb seq. alignées

Choix de 5 références pour générer un multifasta : (source PartitionCaptureMitoParLot.xlsx):
Salmo trutta trutta NC_010007.1
Hucho hucho mitochondrion NC_025589.1
Coregonus muksun mitochondrion NC_028593.1
Coregonus ussuriensis mitochondrion NC_025648.1
Coregonus chadary mitochondrion NC_030175.1
(puis ajout salmo salar dans la suite des tests)

Alignement sur 5 échantillons moins riches en ADN :

lib	[DNA] ng/ul	endo %
L9-13	15,5	1,1
L2-4	9,6	0,9
L5-6	29,8	0,8
L9-4	46,9	0,8
L9-14	27,1	0,8
L5-11	26,6	0,6
L9-6	31,3	0,6

Nb. align. seq. (1)

Sample	NC_010007.1	NC_025589.1	NC_025648.1	NC_028593.1	NC_030175.1	NC_001960.1
L9_14_SH1979-4-6843-a_S56	5687	226	100	75	74	503
L5_6_MB8_S31	1403	602	252	272	241	19719
L5_11_MD14_S36	97040	3487	1799	1723	1802	9437
L9_6_SH1979-4-6843-b_S53	21848	605	322	302	293	1617
L2_4_CD2_S74	15	3	12	1	2	15

(1) Nombre de séquences alignées pour chaque référence (samtools idxstats troisième colonne)

→ **Pas de corrélation entre le nombre de séquences et la quantité d'ADN.**

Alignement sur multiFasta

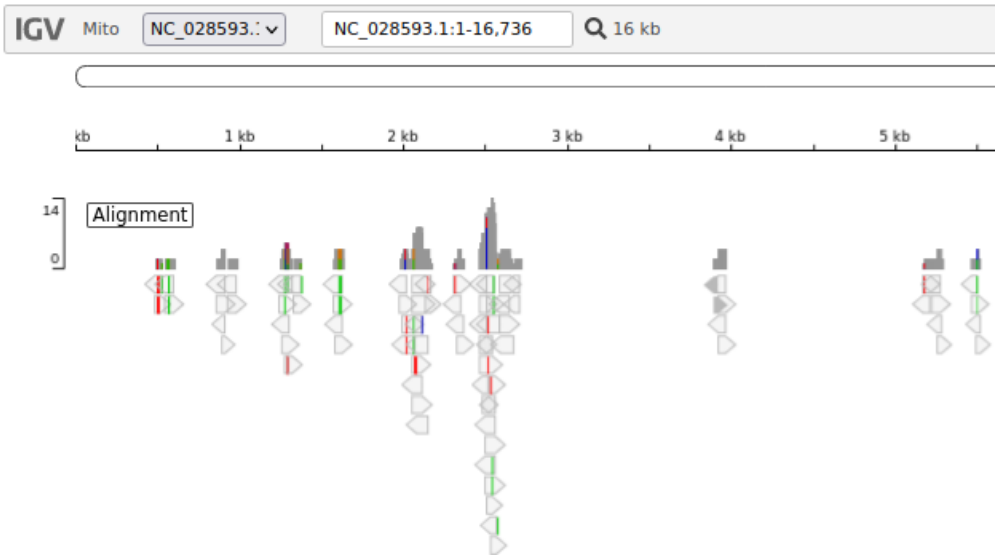
Nb. align. seq. (1)

Sample	NC_010007.1	NC_025589.1	NC_025648.1	NC_028593.1	NC_030175.1	NC_001960.1
L9_14_SH1979-4-6843-a_S56	5687	226	100	75	74	503

```
bash-4.4$ samtools idxstats L9_14_SH
NC_010007.1     16677   5687    91
NC_025589.1     16751   226     3
NC_025648.1     16737   100     0
NC_028593.1     16736   75      2
NC_030175.1     16739   74      0
NC_001960.1     16665   503     5
```

Multi align L9_14_SH1979-4-6843-a_S56

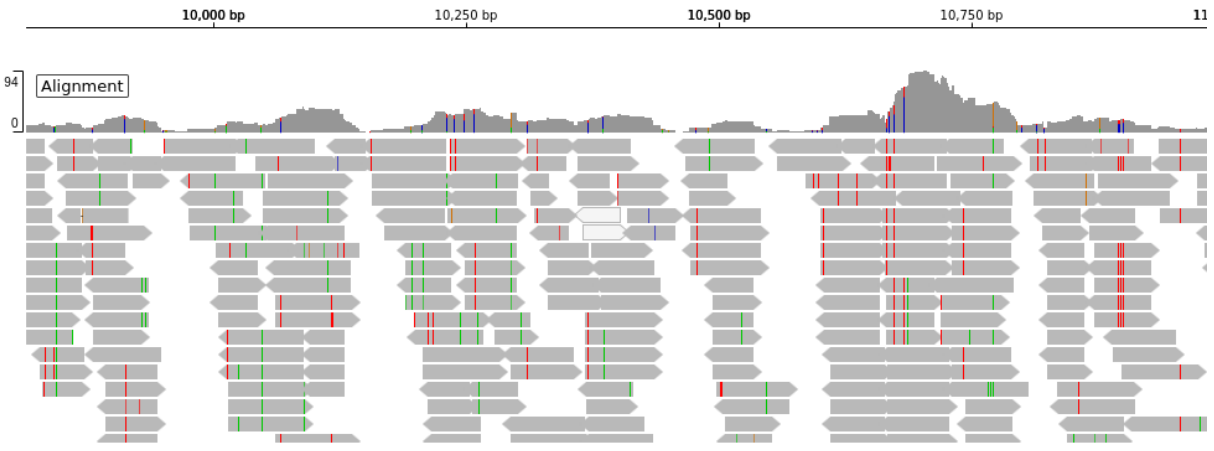
Multi align L9_14_SH1979-4-6843-a_S56



Alignement sur multiFasta

Zoom sur zones moins couvertes :

Zoom sur zones coloriées dans histogramme → Consensus ok



Alignement & Sélection

1/ Alignement de l'ensemble des échantillons disponibles avec le multifasta pour récupérer le nb. De séquences alignées et les profondeurs :

PaleoFish_mito_V2
.ods

2/ Critères de sélection des 27 échantillons :

- 2/3 échantillons par période distincte
- avec une couverture médiane variable : haute - moyenne – basse

Id	Period	sample	mean coverage	median coverage	Ref principale en fonction du nb seq. alignées
L6-11	4400-4000BC	L6_11_SH2000-99-394-g_S46	340,431	630,5	Salmo trutta NC_010007.1 (137754)
L9-15	4400-4000BC	L9_15_SH2000-99-390-c_S57	147,295	223	Salmo trutta NC_010007.1 (44148)
L6-6	4900-4500BC	L6_6_SH2001-106_11-b_S44	32,1258	27	Salmo trutta NC_010007.1 (9938)
L6-4	800-1000 AD =VIKING PERIOD	L6_4_SH1979-4-6843-c_S42	179,328	324,5	Salmo trutta NC_010007.1 (67175)
L6-3	800-1000 AD =VIKING PERIOD	L6_3_SH1979-4-6843-e_S41	132,26	165	Salmo trutta NC_010007.1 (41909)
L9-14	800-1000 AD =VIKING PERIOD	L9_14_SH1979-4-6843-a_S56	20,7994	11	Salmo trutta NC_010007.1 (5687)
L5-6	azilien	L5_6_MB8_S31	59,6934	102	Salmo salar NC_001960.1 (19719)
L4-4	azilien	L4_4_MB9_S14	147,054	45	Salmo salar NC_001960.1 (59671)
L11-6	Late Neolithic (c. 2800 BC)	L11_6_SCO_116_S69	175,585	128	Salmo trutta NC_010007.1 (49085)
L11-3	Late Neolithic (c. 2800 BC)	L11_3_SCO_114_S68	89,1935	47	Salmo trutta NC_010007.1 (24367)
L10-11	Later medieval and post medieval (1400-1800 AD)	L10_11_SCO_1338_S63	499,582	577	Salmo salar NC_001960.1 (166059)
L11-7	Later medieval and post medieval (1400-1800 AD)	L11_7_SCO_1354_S70	291,815	123	Hucho hucho NC_025589.1(10302)
L5-1	magdalenien = paléolithique supérieur	L5_1_MD15_S26	360,818	615	Salmo salar NC_001960.1 (177582)
L5-12	magdalenien = paléolithique supérieur	L5_12_MD9_S37	191,109	333	Salmo salar NC_001960.1 (69810)
L5-2	magdalenien = paléolithique supérieur	L5_2_MD7_S27	79,2106	16	Salmo salar NC_001960.1 (27828)
L2-10	Médiéval	L2_10_MPHB2_S78	397,896	725	Salmo salar NC_001960.1 (139394)
L2-11	Médiéval	L2_11_MPHB3_S79	251,258	341,5	Salmo salar NC_001960.1 (81468)
L10-4	Moderne (XVIe siècle)	L10_4_BC8_S61	528,619	834	Salmo trutta NC_010007.1 189670()
L10-1	Moderne (XVIe siècle)	L10_1_BC5_S58	311,279	416,5	Salmo trutta NC_010007.1 (114132)
L10-5	Moderne (XVIe siècle)	L10_5_BC9_S62	117,123	116	Salmo trutta NC_010007.1 (38431)
L2-7	Néolithique	L2_7_TP2_S76	859,351	1825	Salmo trutta NC_010007.1 (302730)
L2-6	Néolithique	L2_6_TP1_S75	527,888	496,5	Salmo trutta NC_010007.1 (196145)
L3-5	Paléolithique supérieur (Azilien)	L3_5_SM3_S2	311,605	526,5	Salmo trutta NC_010007.1 (96592)
L9-4	Paléolithique supérieur (Azilien)	L9_4_SM1_S52	10,1096	8	Salmo salar NC_001960.1 (3186)
L0-4	Paléolithique supérieur (Magdalénien)	L0_4_EG9_S84	55,5137	127,5	Salmo salar NC_001960.1 (7607)
L0-6	Paléolithique supérieur (Magdalénien)	L0_6_EG13_S86	167,66	50	Salmo trutta NC_010007.1 (71382)
L2-3	Paléolithique supérieur (Magdalénien)	L2_3_EG3_S73	80,2654	8	Salmo salar NC_001960.1 (30140)

Alignement sur référence majoritaire

- Second l'alignement que sur la référence majoritaire de chaque échantillon → pour éviter de dépeupler des zones qui sont partagées entre plusieurs références.
- Ajout d'un filtre sur la qualité des alignements : -q 20
- Profondeur min assez grande pour avoir confiance dans la référence.
- Comparaison des profondeurs min et max :
  - au 1er alignement : sur multifasta et non filtré
  - vs 2nd alignement : sur référence majoritaire et bam filtré q 20

  → **Ceci nous permet donc effectivement de récupérer de la profondeur**

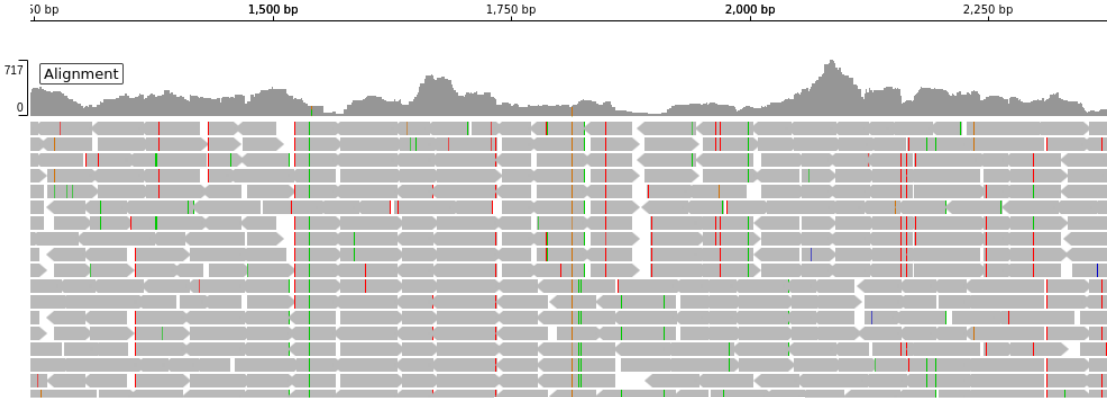
- Explorations manuelles :
  http://genoweb.toulouse.inra.fr/~sigenae/sarah/mito/Second-pass-align/L0_4.igvjs.html

sample	Ref. maj.	1st align. min	1st align. max	2nd align. min	2nd align. max	IGV
L0_4	NC_001960.1	0	571	0	930	IGV
L0_6	NC_010007.1	2	1103	2	1120	IGV
L10_1 *	NC_010007.1	25	1035	57	1052	IGV
L10_11 *	NC_001960.1	1	1536	28	1537	IGV
L10_4	NC_010007.1	43	1643	173	1642	IGV
L10_5	NC_010007.1	6	389	22	407	IGV
L11_3	NC_010007.1	3	248	13	289	IGV
L11_6 *	NC_010007.1	1	737	1	737	IGV
L11_7	NC_025589.1	0	2172	1	4418	IGV
L2_10	NC_001960.1	1	1115	35	1115	IGV
L2_11 *	NC_001960.1	1	689	18	748	IGV
L2_3	NC_001960.1	1	293	6	293	IGV
L2_6	NC_010007.1	51	1996	92	2006	IGV
L2_7	NC_010007.1	119	2626	283	2630	IGV
L3_5	NC_010007.1	24	1044	93	1048	IGV
L4_4	NC_001960.1	2	1274	22	1274	IGV
L5_1 *	NC_001960.1	1	1520	31	1527	IGV
L5_12	NC_001960.1	1	613	18	621	IGV
L5_2	NC_001960.1	1	246	14	263	IGV
L5_6	NC_001960.1	1	301	1	305	IGV
L6_11 *	NC_010007.1	19	1631	103	1635	IGV
L6_3	NC_010007.1	8	462	25	462	IGV
L6_4	NC_010007.1	17	867	42	867	IGV
L6_6 *	NC_010007.1	1	148	1	157	IGV
L9_14	NC_010007.1	1	104	1	136	IGV
L9_15	NC_010007.1	12	437	27	437	IGV
L9_4 *	NC_001960.1	0	54	*	*	IGV

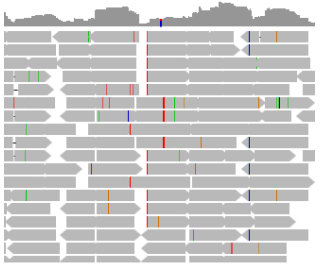
Alignement sur référence majoritaire
IGV L0_6

http://genoweb.toulouse.inra.fr/~sigenae/sarah/mito/Second-pass-align/L0_6.igvjs.html

Zoom sur zones moins couvertes = zones dégradées:
- Même si trou => Blocs de N. Tant que la profondeur n'est pas nulle alors la référence est bonne.
- Même si dégradé -> consensus possible.
- Hétérozygotie, mais pas trop de variations différentes dans les différentes séquences => consensus propre.



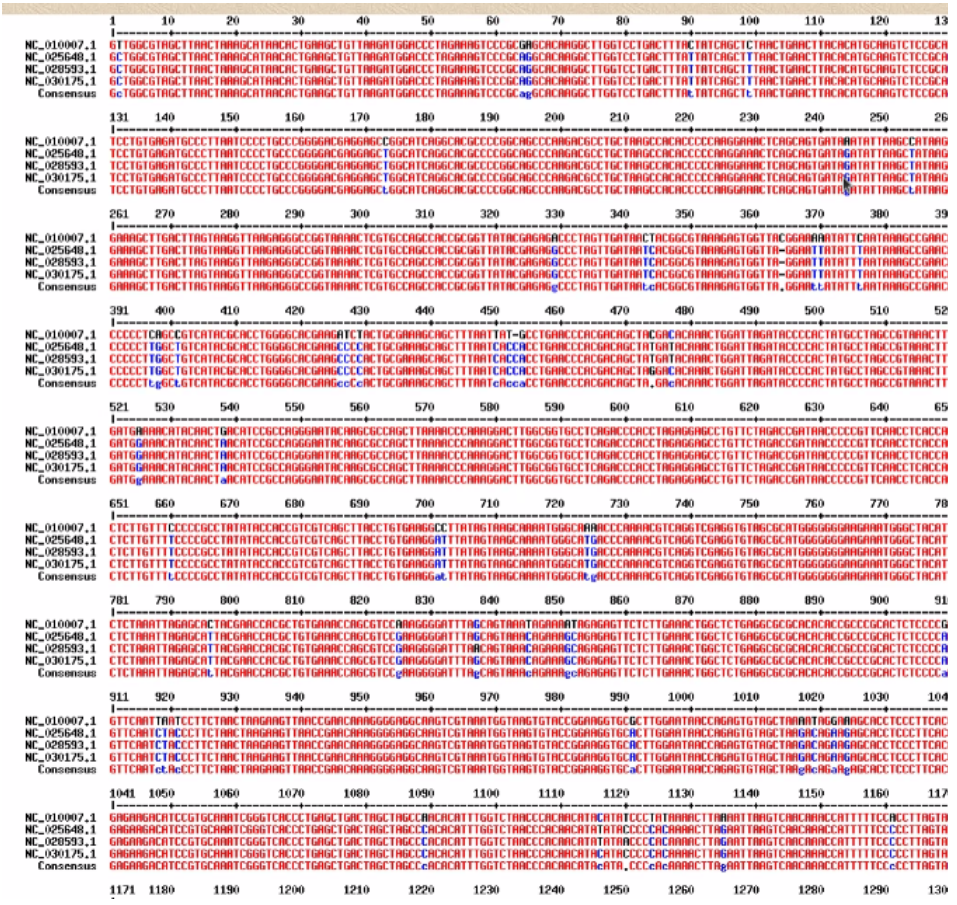
Zoom sur quelques petites variations aléatoires qui ne changent pas le référence :



Perspectives

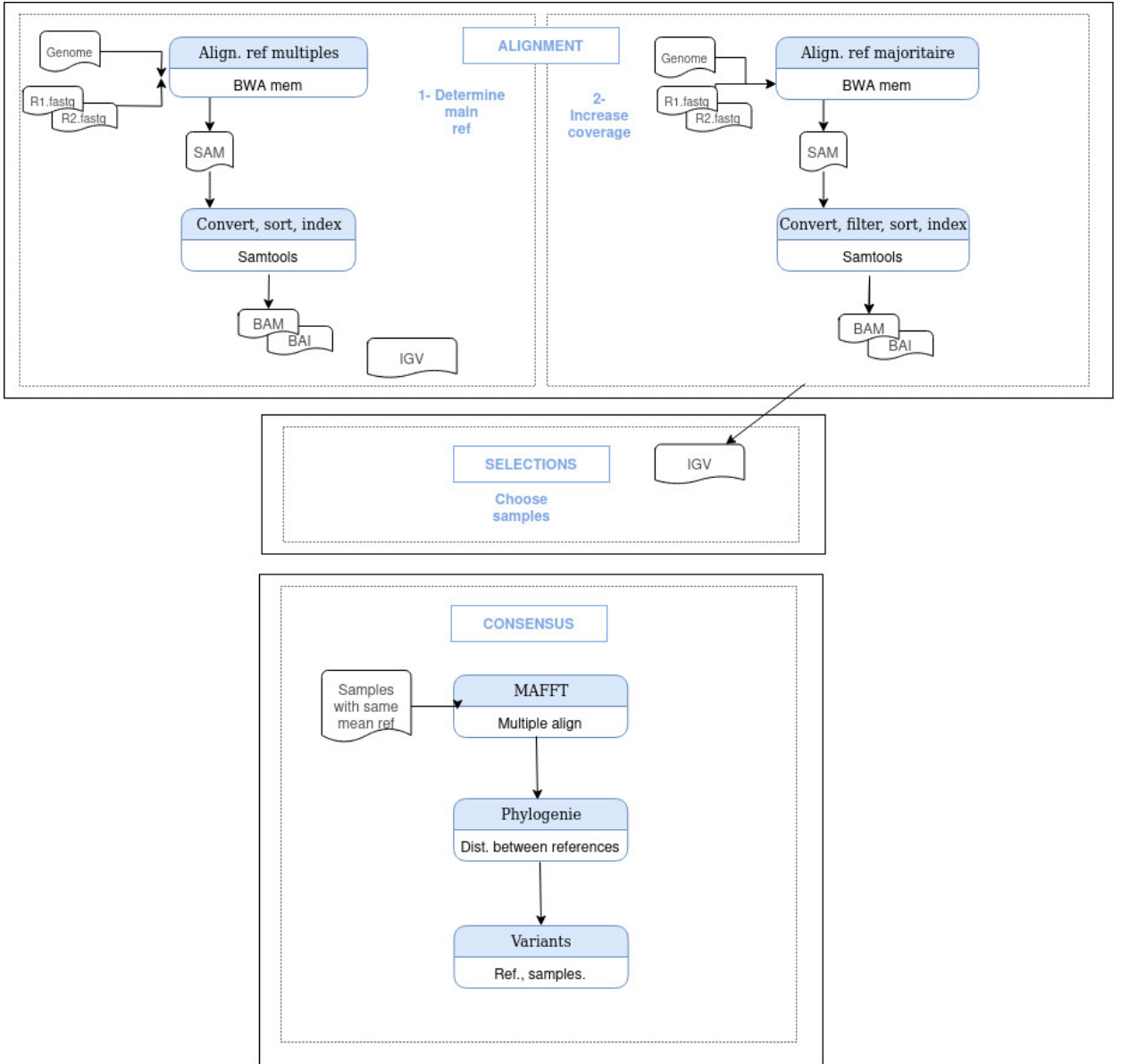
Démarche pour générer les nouvelles référence:

- Contrôler la qualité des alignements avec IGV.
- Sélectionner les échantillons à conserver (avec Joëlle).
- Extraire la séquence consensus de référence en sélectionnant un groupe d'échantillons ayant la même référence majoritaire. Par exemple, les 8 échantillons truite.

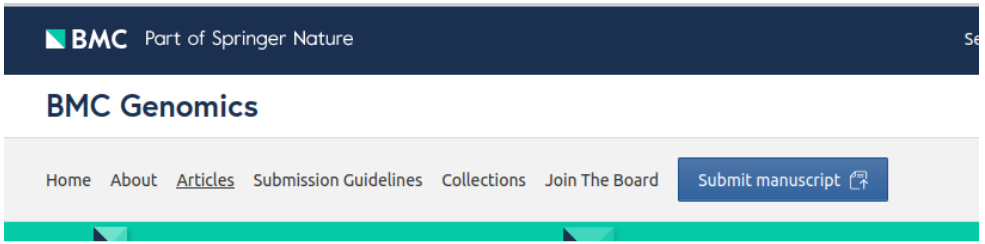


Perspectives

- Générer une nouvelle référence consensus avec les seconds alignements sur ces échantillons avec un outil d’alignement multiple comme MAFFT.
- Analyse phylogénétique (dendrogramme) pour calculer la distance entre les références générées.
- Analyse des variants entre les références et replacer les individus par rapport aux variants.


ALIGNMENT
Genome
R1.fastq
R2.fastq
Align. ref multiples
BWA mem
SAM
Convert, sort, index
Samtools
BAM
BAI
IGV
1- Determine main ref
2- Increase coverage
Genome
R1.fastq
R2.fastq
Align. ref majoritaire
BWA mem
SAM
Convert, filter, sort, index
Samtools
BAM
BAI
SELECTIONS
Choose samples
IGV
CONSENSUS
Samples with same mean ref
MAFFT
Multiple align
Phylogenie
Dist. between references
Variants
Ref., samples.

Perspectives



BMC Part of Springer Nature

BMC Genomics

Home About Articles Submission Guidelines Collections Join The Board Submit manuscript

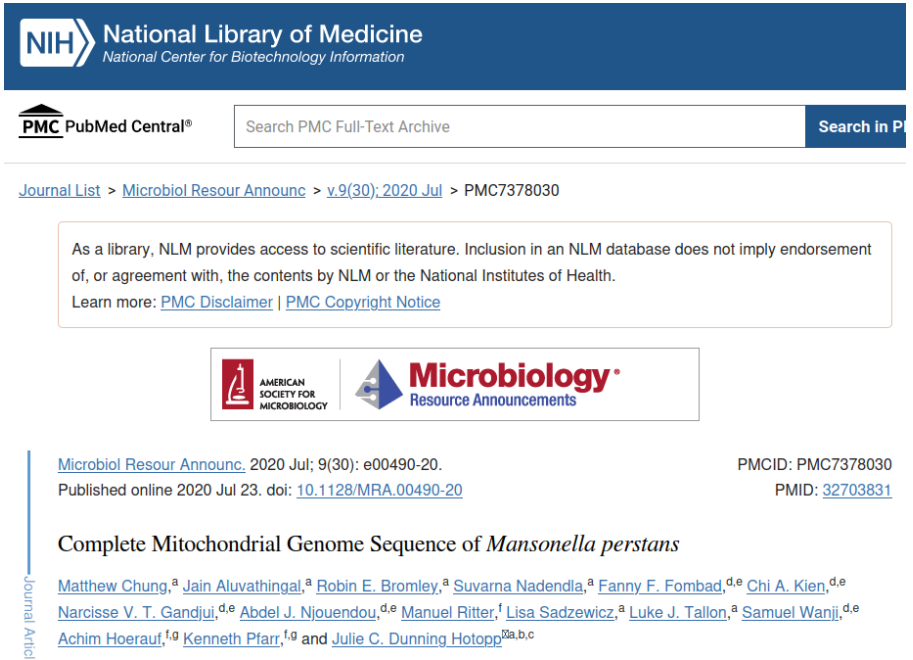
Research | Open access | Published: 23 November 2023

Assembly and analysis of the complete mitochondrial genome of *Forsythia suspensa* (Thunb.) Vahl

Yun Song, Xiaorong Du, Aoxuan Li, Amei Fan, Longjiao He, Zhe Sun, Yanbing Niu & Yonggang Qiao

BMC Genomics **24**, Article number: 708 (2023) | Cite this article

545 Accesses | **1** Altmetric | Metrics



NIH National Library of Medicine
National Center for Biotechnology Information

PMC PubMed Central® Search PMC Full-Text Archive Search in PI

Journal List > Microbiol Resour Announc > v.9(30); 2020 Jul > PMC7378030

As a library, NLM provides access to scientific literature. Inclusion in an NLM database does not imply endorsement of, or agreement with, the contents by NLM or the National Institutes of Health.
Learn more: PMC Disclaimer | PMC Copyright Notice

American Society for Microbiology | Microbiology Resource Announcements

Microbiol Resour Announc. 2020 Jul; 9(30): e00490-20. PMCID: PMC7378030
Published online 2020 Jul 23. doi: 10.1128/MRA.00490-20 PMID: 32703831

Complete Mitochondrial Genome Sequence of *Mansonella perstans*

Matthew Chung,[a] Jain Aluvathingal,[a] Robin E. Bromley,[a] Suvarna Nadendla,[a] Fanny F. Fombad,[d,e] Chi A. Kien,[d,e] Narcisse V. T. Gandjui,[d,e] Abdel J. Njouendou,[d,e] Manuel Ritter,[f] Lisa Sadzewicz,[a] Luke J. Tallon,[a] Samuel Wanji,[d,e] Achim Hoerauf,[f,g] Kenneth Pfarr,[f,g] and Julie C. Dunning Hotopp[a,b,c]

Journal Articl

Conclusion

- Grandes variabilités sur les quantités de séquences produites.
- Certains échantillons corrompus ou absents ~ 30 perdus / 89.
- Beau potentiel avec ~ 20-30 truites et ~20 saumon.
- Voir si des zones géographiques et des périodes historiques sont perdues.